Confidence-Aware Learning for Person Re-identification with Noisy Labels

Duhyun Kim[1] and Jae-Young Sim[1,2,†]
[1]Graduate School of Artificial Intelligence, [2]Department of Electrical Engineering,
Ulsan National Institute of Science and Technology, Republic of Korea
E-mail: {duhyunkim, jysim}@unist.ac.kr

***Abstract*—Accurately labeling a large amount of data for person re-identification is a significant challenge. In this paper, we introduce a technique to effectively perform person re-identification even in the dataset with noisy labels. Leveraging the widely observed phenomenon that data with wrong labels tends to have large loss values, we fit the Gaussian mixture model(GMM) to estimate confidence which is the probability of the sample being noise-labeled. We propose confidence-aware learning that appropriately reflects confidence to balance between mitigating the impact of samples with noisy labels and guiding anchors to the complete positive and negative samples. Additionally, we refine the GMM to enhance the accuracy of confidence for each data sample even in a lack of data situation. Experimental results demonstrate that our methods are effective techniques for handling noisy labels in person re-identification.**

I. Introduction

Person re-identification (Re-ID) aims to identify the same person across various viewpoints obtained from different cameras. This technology has progressed because it is crucial in surveillance for safety and lifelogging. Existing supervised Re-ID performed well using many person image datasets and high-quality identity labels during training. However, accurately labeling large amounts of frames is challenging due to the minor variations in inter-class such as appearance, and significant variations in intra-class such as viewpoints and resolution. In addition, establishing guidelines for perfect labeling and hiring personnel with expertise in labeling can be time-consuming and expensive. Therefore, even if there is noise in the labeled data, low-cost labeling methods like crowdsourcing or annotation tools can still be utilized. Unfortunately, when trained with noisy labels, various supervised methods suffer from performance degradation as the model learns to consider different individuals as the same person.

Various research efforts are emerging to address the impact of noisy labels, but applying these directly to Re-ID is hard because of the large number of ID classes and severe class imbalance. Fig. 1 shows the comparison of the amount of data within the classes. In the classification dataset [1][2], all classes have sufficient data and a low degree of class imbalance. In contrast, Re-ID datasets typically exhibit severe class imbalance, making distinguishing noise within ID classes difficult with tale class data. Moreover, the large number of ID classes in person re-identification datasets complicates the identification of a sample's correct ID.

Existing studies on learning with noisy labels [3][4][5] did not handle these issues. They commonly utilize predicted ID labels during training by adopting a soft label strategy. In other words, they believe that using predicted ID labels is more reliable than relying on noisy labels. However, Re-ID datasets have many ID classes, making it even more challenging to accurately predict IDs during training. Consequently, using predicted ID classes may introduce further inaccuracies that interfere with model training. Therefore, we propose a new learning strategy that does not rely on a soft label strategy.

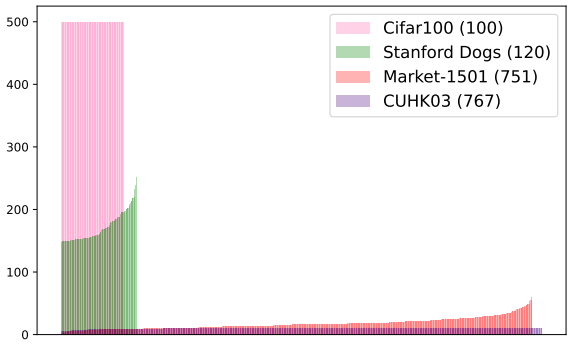

Fig. 1. Comparison of the amount of data within the class by four datasets. Cifar100 [1] for classification, Stanford Dogs [2] for fine-grained classification, Market-1501 [7] and CUHK03 [8] for re-identification. The number in the legend means the number of classes in the dataset.

Additionally, the data imbalance issue implies that the likelihood of noise is different according to ID class. Individuals from frequently appearing IDs in the dataset will have more variance and, consequently, a higher probability of having noise labels. This highlights the need to adequately consider whether a sample within a class is noisy or not. However, image recognition studies [5][3][6] assume that there is the same probability of the noise for each class, so they do not handle the issue of inter-class data imbalance. To tackle these challenges, we exploit the observation that samples with noisy labels lead to higher loss values compared to those with clean labels. We utilize a two-component Gaussian Mixture Model (GMM) for loss values to assess the label confidence of each instance, based on the probability that a sample's loss value belongs to a clean distribution component. In this phase, we devise GMM refinement to obtain more reliable confidence for each sample. This is crucial because GMM modeling tends to perform poorly when there is insufficient data within

the ID class. After that, we incorporate label confidence into model training by preventing the model from being trained on incorrectly labeled samples when access to the correct ID class is not available during training. The main contributions of this paper are as follows.

- In this paper, we devise the GMM refinement to get more reliable confidence for the label considering the inter-class data imbalance.
- We propose a new learning method to reflect the confidence of labels for each instance by reducing the impact of incorrect labels and helping the model optimize by guiding input to a completely positive and negative sample.
- We demonstrate our proposed method effectively handles the issue caused by noisy labels on Marekt-1501 and CUHK03.

II. Related Works

A. Person Re-ID

Many supervised methods achieve the highest performance thanks to datasets with clean labels. TransReID [9] proposed a transformer-based strong baseline and a jigsaw patch module to handle occlusion and misalignment issues. CDNet [10] introduced combined blocks to extract features more suitable for retrieval by constructing two branches with different kernel sizes. However, these supervised methods rely on the correct labels with expansive labeling cost and show performance degradation under the noisy labels setting.

B. Image Classification with noisy labels

Co-teaching [11] trained the model with the small loss value first and prevented overfitting by exchanging results between two identical models. JoCoR [12] also adopted a two-identical model strategy and trained the model only when the predictions of the two models were the same. TCL [6] employed a GMM for features to obtain reliability scores and utilized contrastive learning, assuming that the two samples are always considered positive when creating mixup views. SNSCL [5] proposed a solution to address the issue of noisy labels in fine-grained classification, where the data for each class exist in a closer feature space. They estimated the reliability score of the data using a two-component GMM of loss values and leveraged a soft label strategy between the model's predictions and the ground truth. However, these methods also use a soft-label strategy without concerning the large number of ID classes. And they do not handle the issue of severe class imbalance because they assume that the number of noisy labels in all classes is similar.

C. Person Re-ID with noisy labels

There are existing Re-ID methods designed to address the challenges of noisy labels. DNet [13] assigned uncertainty into features to prevent the model from optimizing by noise samples. PurifyNet [4] trained model with soft label strategy using predicted ID classes and introduced a technique to assign lower weights to samples located away from the center of each class. CORE [3] introduced a co-refining strategy in which two identical structures collaboratively refine each other using soft label cross entropy by exchanging their predicted results during the training. TSNT [14] also adopted a co-refining strategy and introduced a more suitable triplet loss to handle noisy labels by selecting more reliable hard-positive and negative samples. [4][3][14] utilize a soft label strategy assuming that predicted ID class during the training is more reliable than noisy labels. However, it is hard to get the exact correct ID class during the training because there are a lot of ID classes in the re-identification. LRP [15] evaluates label confidence by determining how many close samples share the same ID class. LRNI-HSR [16] refines labels before training by relying on the judgments of a pretrained network. ICLR [17] refined label samples when they were far from the ID's center feature and adjusted the training contribution to give higher attention to samples estimated to be noisy in the early epochs.

III. Proposed Methods

We reflect the confidence of the ID label to minimize the model's performance degradation caused by incorrect labels. This process is composed of two steps and is repeated for every epoch. First, we estimate the sample's confidence, representing the probability of being the correct label, with GMM refinement to get a more reliable GMM even when there is a lack of data within the class. Second, the model should be trained with the Re-ID loss to suppress low-confidence samples and Instance loss to make low-confidence samples closer to the complete positive samples and far from negative samples. Fig. 2 illustrates these two steps, where the blue dashed line represents the confidence estimation step with frozen model weights, and the red line represents confidence-aware learning.

A. Confidence Estimation

We employ a two-component Gaussian Mixture Model(GMM) for each class to get the probability of a sample having the correct labels [5]. This is a common approach based on the observation that samples with noisy labels tend to have higher loss values than samples with clean labels. As Fig. 3, this knowledge can also adapted to the Re-ID task. We can regard the scale of loss as a clue to identifying samples with a noise label. After fitting the GMM, we estimate the posterior probability of belonging to the clean distribution of all the person instances in the training dataset, denoting confidence γ. The confidence is renewed before every training epoch.

However, because of the data imbalance in the Re-ID dataset, all the ID classes do not always have sufficient samples with noisy labels to get reliable GMM. For example, if one class is composed of only samples with correct labels, the second component, which should represent the distribution of noise loss values, encounters an issue where it is composed of clean samples rather than noisy samples. Furthermore, when the number of noise samples is extremely small, proper modeling may not occur because the noise samples do not necessarily have larger loss values than clean samples. To

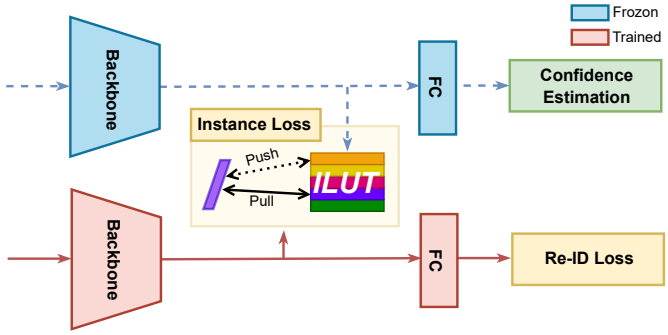

Fig. 2. Overview of our proposed methods.

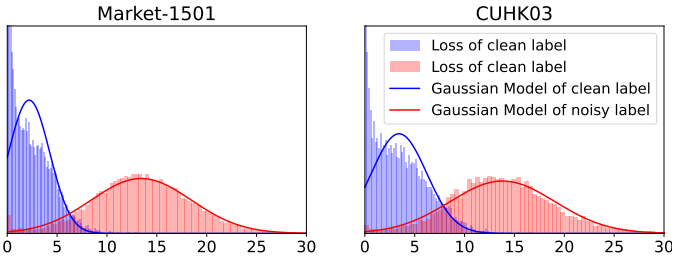

Fig. 3. Histogram of loss values calculated by correct label and incorrect label, and Gaussian distribution model with loss values in two datasets Market-1501 and CUHK03.

address this, we devised GMM refinement (GR), which is the trick to getting a more reliable GMM for our re-ID situation. When modeling the GMM for a specific ID class, we utilize the loss values of samples within the ID class and samples randomly selected from another class, representing the role of fake samples with noisy labels. As a result, the second component of the GMM can always represent the distribution of the noise loss value.

B. Confidence-Aware Learning

Cross-entropy loss and triplet loss are used to train the Re-ID model generally. However, they are ineffective in noisy label settings because of their strong dependence on the ID labels. To address this issue, we introduce a learning strategy for generating robust models from noisy labels using estimated confidence.

1) Re-ID Loss: If we train the model with the general cross-entropy loss in a noisy situation, samples with noisy labels are guided to the wrong ID class, harming the model training. To tackle this issue, we adaptively adjust training loss, ensuring that samples with low confidence exert a lesser impact on the model compared to samples with high confidence by

$$\mathcal{L}_{\mathrm{id},i} = -\gamma_i \log \frac{\exp(\boldsymbol{x}_{i,y_i})}{\sum_{j=1}^{K} \exp(\boldsymbol{x}_{i,j})}, \tag{1}$$

where γ_i is the confidence of i-th input, $\boldsymbol{x}_{i,j}$ is the prediction score for i-th input being in class j, y_i is the ground-truth label of i-th input, K is the number of ID class.

2) Instance Loss: Furthermore, we introduce a new loss function that aims to bring features closer to those in the Instance Look-Up Table (ILUT). The ILUT is maintained based on the same instance number, which refers to the unique number assigned to each frame. We especially store normalized features on ILUT before starting every epoch. Our concept is that a complete positive sample in the dataset is identical to itself. Therefore, the instance loss aims to bring the input features closer to the features obtained from the same instance after the previous epoch. We also regard the farthest features in ILUT from the input's features as complete negative samples. Therefore, we argue that it is a more effective approach that low-confidence samples learn with complete positive and complete negative samples rather than only suppress noisy data by our $\mathcal{L}_{\mathrm{id}}$. In this term, we add the weighting term $1-\gamma$, which means the probability of a sample with a noisy label. So we train a model by

$$\mathcal{L}_{\mathrm{in},i} = w_{\mathrm{in}}(1-\gamma_i)\log(1+\exp(d(\boldsymbol{f}_i, \boldsymbol{g}_i^p) - d(\boldsymbol{f}_i, \boldsymbol{g}_i^n))), \tag{2}$$

where $\boldsymbol{f}_i$ is feature vector of i-th input, $\boldsymbol{g}_i^p$ is i-th feature vector of ILUT, and $\boldsymbol{g}_i^n$ is the features in ILUT most far from the $\boldsymbol{f}_i$, w_{in} is the hyper parameter, $d(\cdot,\cdot)$ means euclidean distance. Using a complete positive and negative sample is considered the most reasonable choice in settings with noisy labels, where accessing the actual correct label is challenging.

IV. Experimental Results

A. Experimental Settings

1) Benchmark Datasets with Noisy labels: We conduct experiments on three benchmark datasets, Market-1501 [7], CUHK03 [8] in various noisy label settings. We generate noisy labels in the same way as previous studies [4][3]. For the random noise setting, we exploit noisy labels by replacing the IDs of some percentages of labeled samples with randomly selected other IDs using a fixed random seed. For the pattern noise setting, which is more similar to real-world conditions, we assign the secondary closest class as noisy labels.

2) Implementation Details: Our experiments are conducted on a single NVIDIA RTX-3090 GPU using PyTorch. We utilize a simple baseline, which is composed of ImageNet pre-trained ResNet50 [18] backbone with two fully connected layers and one batch normalization while training with general cross-entropy loss. The initial learning rate is set to $10-2$, adjusted to $10-3$ at epoch 20, and then decreased to $10-5$ beyond epoch 50. The batch size is set to 32. The weighting parameter w_{id} is set to $10^{-}3$. Data transformations include random crop to 256×128 and random horizontal flip with a probability of 0.5.

We also conduct experiments on the two supervised methods as a strong baseline, TransReID [9] and CDNet [10], by exchanging their all of cross-entropy losses and triplet losses with our $\mathcal{L}_{\mathrm{id}}$ and $\mathcal{L}_{\mathrm{in}}$ respectively. The initial learning rate is set to $10^{-}1$ and decays by multiplying $10^{-}2$ for every 20 epochs. Other settings are the same as their paper.

Method	Random Noise						Pattern Noise			
	10%		20%		30%		10%		20%	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
DNet [13]	82.3	61.5	77.0	53.4	77.1	44.3	52.4	27.0	49.3	24.4
PurifyNet [4]	84.2	64.3	83.1	63.1	81.4	60.2	81.8	63.2	77.8	56.2
CORE [3]	85.5	67.7	84.1	66.2	81.2	60.6	83.2	66.0	81.3	62.4
LRP [15]	88.9	71.4	88.8	70.5	86.3	67.0	88.5	70.7	86.7	67.6
LRNI-HSR [16]	86.9	68.8	84.5	65.6	81.9	62.2	85.3	67.7	84.1	65.6
ICLR [17]	88.0	70.8	87.2	70.3	86.4	69.7	86.8	68.0	83.6	64.9
TSNT [14]	-	-	90.3	76.0	-	-	-	-	90.9	76.4
Proposed	90.4	69.0	90.0	68.3	88.0	61.6	89.0	64.9	86.3	59.3

TABLE I
COMPARISON OF EXISTING METHODS AND OUR PROPOSED METHODS ON MARKET1501 WITH RANDOM NOISE AND PATTERN NOISE.

Method	Market1501		CUHK03	
	R-1	mAP	R-1	mAP
CDNet [10]	86.0	63.1	27.6	27.2
+Proposed	87.4	69.4	29.9	29.4
TransReID [9]	87.9	70.8	45.6	41.6
+Proposed	**91.8**	**79.2**	**50.3**	**47.4**

TABLE II
COMPARISION OF EXISTING METHODS AND OUR PROPOSED METHODS ON STRONG BASELINE [10][9] WITH 20% NOISE LEVEL.

Method	Random Noise						Pattern Noise			
	10%		20%		30%		10%		20%	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
DNet [13]	32.3	31.8	24.3	24.2	16.8	17.4	10.5	10.9	8.8	9.5
PurifyNet [4]	32.8	32.8	30.2	29.2	26.6	26.4	33.6	32.9	29.2	29.2
CORE [3]	40.4	39.6	34.4	35.0	34.6	34.7	38.6	37.2	34.4	33.2
TSNT [14]	-	-	49.4	48.3	-	-	-	-	50.2	49.0
Proposed	41.1	39.1	39.7	37.6	31.0	30.5	36.6	33.4	30.6	28.1

TABLE III
COMPARISON OF EXISTING METHODS AND OUR PROPOSED METHODS ON CUHK03 WITH RANDOM NOISE AND PATTERN NOISE.

Method			Market1501		CUHK03	
$\mathcal{L}_{\text{id}}$	$\mathcal{L}_{\text{in}}$	GR	R-1	mAP	R-1	mAP
			86.1	56.7	23.4	23.3
✓			88.7	64.9	34.4	32.6
✓	✓		88.6	63.7	34.4	32.8
✓		✓	88.8	66.8	32.1	31.9
✓	✓	✓	**90.0**	**68.3**	**39.7**	**37.6**

TABLE IV
ABLATION STUDY OF OUR PROPOSED METHODS WITH 20% NOISE LEVEL.

3) Evaluation Metrics: To evaluate the model's Re-ID performance, we utilize the Rank-k(R-k) metric and Mean Average Precision(mAP). R-k indicates how many nearest k persons are in the true positive case. The mAP metric represents the mean of average precision, denoting the ratio of true positives among the model predictions.

B. Comparison with Robust Re-ID with Noisy Labels

TABLE I, TABLE III illustrates that our approach is competitive compared to other state-of-the-art methods on two datasets, Market1501 and CUHK03. In the table, red indicates the highest performance, while blue indicates the second highest performance. Some mAP performances are lower than those of existing methods compared to the Rank 1 score relatively. Our $\mathcal{L}_{\text{in}}$ learns to make i-th feature behave similarly to the information stored in i-th feature in ILUT during training. As a result, if the query and gallery images have the same viewpoint or pose, our loss effectively improves the Rank-1 score. In terms of mAP, it may lead to incorrect matching results because the model can be fitted to specific viewpoints and poses. However, we argue that treating each sample as a complete positive sample is the best choice under the noisy label setting.

C. Flexibility of Usability

We also apply our proposed methods to state-of-the-art supervised Re-ID networks, such as those presented in [9] and [10]. Despite the strong performance of these networks, ID noise degrades their performance. However, our proposed method significantly improves performance, as shown in TABLE II. Some existing methods that utilize the co-teaching [3][14] approach have the drawback of doubling the parameters when applied to these supervised techniques. However, the proposed method has the advantage of not requiring any additional parameters. Additionally, using TransReID along with the proposed method outperforms existing methods in TABLE I and TABLE III. It is important to note that our method does not require additional parameters, making it easy to apply to any re-identification networks.

D. Ablation Study

1) Quantitative Results: Table IV is the ablation study of each component of our proposed methods. First of all, our $\mathcal{L}_{\text{id}}$ suppress low-confidence samples when model training, so it is effective under the data containing some incorrect labels. Furthermore, we conducted experiments to demonstrate the effectiveness of the $\mathcal{L}_{\text{in}}$, refer to the third low of the table. Our confidence-aware learning adjusts the two loss functions by balancing the loss scale adaptively utilizing the probability of the sample having the correct label. This allows the model to be optimized correctly even if the sample has low confidence. GMM refinement(GR) enhances the confidence estimation in situations where few noise samples are included within an ID class. Therefore, it improves the performance of the loss that reflects confidence, denoted as $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{in}}$.

2) GMM Refinement: Our GMM refinement overcomes situations where the intra-class has significantly different proportions of correct labels and noisy labels. When there are no noisy samples within a class, a two-component GMM is created based on the loss values of clean samples. This implies a tendency that outlier samples within the intra-class will have lower confidence values, leading to GMM's second component representing clean sample loss values. In Fig. 4, the third sample among the women wearing white t-shirts has a dark background, making it seem different from other samples. Before applying the proposed technique, this clean sample has a low value of 0.02 despite having a correct label. However, after applying GMM refinement as shown on the right, it can obtain a high confidence value of 0.95, which

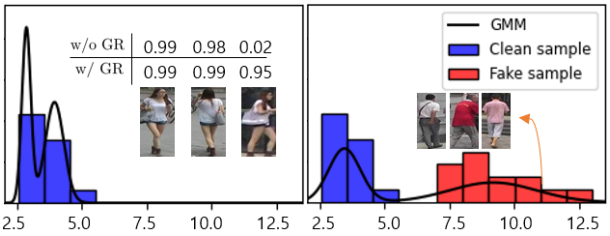

Fig. 4. Distribution of the loss values and visualize some samples with a correct label in the same ID class with their estimated confidence on Market1501.

can correctly influence the training process by our confidence-aware learning.

V. Conclusion

In this paper, we propose a universal technique that can be adapted to any re-identification network without additional learning parameters. To alleviate the confusion caused by noisy labels, we reflect the sample's confidence to model training with our confidence-aware learning. We also devise the GMM refinement to get a more reliable model for loss values even if the intra-class has few samples for noise distribution. Experimental results show that our proposed method, when applied to a strong baseline, outperforms existing methods that learn with noisy labels.

References

[1] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," *Toronto, ON, Canada*, 2009.

[2] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, 2011, pp. 806–813.

[3] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Trans. Image Process.*, vol. 31, pp. 379–391, 2021.

[4] M. Ye and P. C. Yuen, "Purifynet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.

[5] Q. Wei, L. Feng, H. Sun, R. Wang, C. Guo, and Y. Yin, "Fine-grained classification with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 11 651–11 660.

[6] Z. Huang, J. Zhang, and H. Shan, "Twin contrastive learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 11 661–11 670.

[7] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1116–1124.

[8] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 152–159.

[9] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15 013–15 022.

[10] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2021, pp. 6729–6738.

[11] B. Han, Q. Yao, X. Yu, *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8536–8546.

[12] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 726–13 735.

[13] T. Yu, D. Li, Y. Yang, T. M. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 552–561.

[14] M. Liu, F. Wang, X. Wang, Y. Wang, and A. K. Roy-Chowdhury, "A two-stage noise-tolerant paradigm for label corrupted person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, pp. 4944–4956, 2024.

[15] Y. Chen, M. Liu, X. Wang, F. Wang, A.-A. Liu, and Y. Wang, "Refining noisy labels with label reliability perception for person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 9479–9490, 2023.

[16] X. Zhong, S. Su, W. Liu, X. Jia, W. Huang, and M. Wang, "Neighborhood information-based label refinement for person re-identification with label noise," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, 2023, pp. 1–5.

[17] X. Zhong, X. Han, X. Jia, *et al.*, "Iclr: Instance credibility-based label refinement for label noisy person re-identification," *Pattern Recog.*, vol. 148, p. 110 168, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.